Document comparison done by DeltaView on Friday, October 22, 2004 3:16:50 PM

| Input: | |
|---|---|
| Document 1 | iManageDeskSite://washdms1/WASH/1194740/1 |
| Document 2 | iManageDeskSite://washdms1/WASH/1282409/1 |
| Rendering set | Standard |

| Legend: | |
|---|---|
| Insertion | |
| Deletion | |
| Moved from | |
| Moved to | |
| Style change | |
| Format change | |
| Moved deletion | |
| Inserted cell | |
| Deleted cell | |
| Moved cell | |
| Split/Merged cell | |
| Padding cell | |

| Statistics: | |
|---|---|
| | Count |
| Insertions | 77 |
| Deletions | 136 |
| Moved from | 4 |
| Moved to | 4 |
| Style change | 0 |
| Format changed | 0 |
| Total changes | 221 |
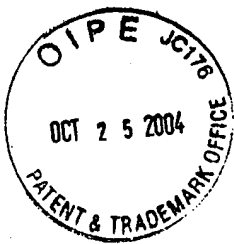
# SPECIFICATION KNOWLEDGE DISCOVERY APPARATUS AND METHOD

[Electronic Version 1.2.8]

Knowledge Discovery Apparatus and Method

Federal Research Statement

This invention was developed entirely under the internal efforts of EagleForce Associates, Inc. ("EagleForce"). EagleForce claims all rights with regard to this patent.

Background of Invention

[0001]           1. Field of the Invention

# BACKGROUND

[0001]    [0002] ThisThe invention is directed to an apparatus and method for performing knowledge discovery by extracting elements of information that are useable to an analyst with regard to an area of inquiry, whether or not that inquiry has been formally framed or the "inquiry" is generated by the apparatus in the course of automated processes.

[0003]           2. Description of the Related Art

[0002]    [0004] There are many applications performing Knowledge Discovery (KD), ranging from Federal and Defense intelligence to business intelligence. [0005] ————Often, in such applications, many KD tools are used to perform specific steps in the KD process as identified in Claims (1) through(7). More recently, various suites of such tools have been assembled to perform sequences of related KD operations. An example of such is the architecture adopted for the (2002) Joint Intelligence Virtual Architecture system. These systems are limited by the lack of either a Feedback Loop or a Utility Function modifying the

Feedback Loop.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0003] FIG. 1 Illustrates the challenge of scalability, which shows how very large data corpora must be processed in order for to extract meaning relative to a given inquiry.

[0004] FIG. 2 is exemplary schematic views of the seven levels for a complete KD architecture includes five representation levels (1 through 5) and two control levels (6 and 7), in accordance with the invented method and apparatus. This figure shows the EagleForce "Representation Levels" concept, which is a foundation for building a knowledge discovery architecture. Levels 1 through 5 are detailed with Level 0 indexing (not shown) being reserved for the ingestion of extremely large data sets. Level 6 provides feedback control of lower levels, and Level 7 contains a utility function that is used to optimize feedback. This scalability serves to significantly enrich the metatagging process.

[0005] FIG. 3 is provides a schematic view of data flow through the apparatus, including the optional step 0, but not reflecting optional step 5c, beginning with the original data corpus and the transformation of the data corpus through the operations performed upon the data corpus.

## DESCRIPTION

[0006] [0006] This invention overcomes the above-noted disadvantages. An apparatus in accordance with this invention is constructed to receive data feeds from one or more data sources, where the data feeds may include live and / or stored data, including "structured" (database) data, unstructured (e.g., document, web page), semi-structured (e.g., military Commander's Intent orders, militaryFrag(mentation) orders, or military or commercial email), along with audio, video, and / or image data. It is the intent of described metataggng

2

methodology and apparatus to provide the highest and best use of the indexing, classification, and categorization of information resident within the collateral networks. The distinguishing feature of the methodology is the use of the "EF Feedback Loop", a process that incorporates the highest and best use of multiple COTS tools. The feedback loop is a widely accepted calibration concept, commonly deployed in this environment for elements of ranking algorithms, type weights, and type proximity-weights. The feedback loop is used in conjunction with one or more of the EF Utility Function(s). The purpose of the utility functions is to iteratively adjust the parameter controls sent back via the feedback loop process in order to maximize results according to a given benefit or utility.

[0007] [0007]-The primary challenges associated with retrospective metadata tagging are:

> [0008]-1. Creating the right metadata "concept classes" that identify those corpus elements (e.g. documents, pages, paragraphs) containing inquiry-relevant concepts, and

> [0009]-2. Ensuring scalability. [0010]———The issue of scalability compels us to use an architectural suite of integrated COTS tools as integral to the apparatus, along with the control mechanisms of feedback loops governed by utility functions. This is the only means by which metadata tagging can be retrospectively done, while still maintaining the ability to handle very large (e.g., order-of-terabyte, or $O(1012)$, sized corpora.

[0008] [0011]-The scalability issue also motivates us to useis dealt with by using an integrated COTS suite to reduce the manpower overhead and minimize the level of human interaction required to support the retrospective markup process, while still maintaining the quality of the metadata markup needed for precision searching.

[0009] [0012] The key issue in controlling scalability, and in reducing manpower overhead, is to determine correct parameter settings governing the metadata tagging process as well as information retrieval in response to metatag-based queries. This is undoubtedly the most significant challenge in the data analysis and metatagging process. One reason that this is so challenging is that when retrospective metadata tagging is introduced as an additional processing stage on top of preliminary data metatagging, the issues associated with corpus size and scalability are exacerbated. Thus, it is crucial to find a method by which metadata tagging can be done, both initially and retrospectively, in a manner that both makes precise inquiry possible and which allows scaling to very large corpora.

[0010] [0013] Google patent holders, Drs. Sergey Brin and Lawrence Page, who in their paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine," state, "Figuring out the right values for these parameters is something of a black art", express the importance of this challenge.

[0011] [0014] Like most others, Drs. Brin and Page place the user as the initial and primary element(s) of the feedback loop. There, the "user may optionally evaluate all of the results that are returned." But it is precisely this positioning that becomes untenable as very large corpora are considered. This "Google" process, common among most COTS tagging and search products, has clearly achieved less than satisfactory results in the challenging intelligence data-parsing environment. Even user-oriented search training functions ultimately only serve to constrain results based on the limitations of a particular tool's mathematical capabilities.

[0012] [0015] To enhance this well-established query process into structured, unstructured, and semi- structured data, many in the Defense, Intelligence, and commercial environment have begun developing suites of tools that utilize different algorithms against the same data set. Two major issues evolve when

4

using such suites:

[0016] 1. Query results using these suites generally differ based on the order of the data flow.

[0017] 2. The results are extremely inconsistent and become virtually unusable as the data corpora expand.

[0013]   [0018] The latter issue of results inconsistency is directly related to the issue of scalability, which is a primary concern when dealing with retrospective metadata tagging. Generally, the metaschema between the tools is unique to the individual product and integration, even that which extends to the API level, allowing the individual tool to read and optimize its portion of the metadata. Knowledge is organized and presented in an extremely robust manner when the data corpora are small. However, as the size of the originating file expands, the discovery of relevant knowledge and entities/concepts to tag, suffers greatly.

[0014]   [0019] The EagleForce approach to minimizing The present invention minimizes the user interaction level required for precise searching is to by first define defining a functional architecture in which different levels of knowledge representation and knowledge processing are used in successive manner. Both initial and retrospective metadata tagging are done at Level 1. Higher levels allow for different degrees of correlation among the data. When these correlations are done, it is possible to generate focused and pertinent retrospective metadata tagging directives. This is done partially through modifying the ranking function that guides metadata tagging. The modified ranking function is used to present the rank impact of the change on all previous searches.

[0015]   [0020] Here the EF FeedBack Loop runs a Level 1 classifier tool at a very simple level as a first pass. This serves to focus on getting those documents that have the highest, richest data relative to the inquiry as we position our classifier to operate with a very tight sigma - i.e., a document has to have lots of hits on

very simple, core keywords in order to be selected and moved forward. For this purpose, we use a Bayesian classifier with Shannon relevance ranking. The value of the EF Feedback Loop and the EF Utility function allows the use of multiple independent or collective Level 1 tools. The EF Feedback Loop and the EF Utility Function apparatus is employed to control the processing limits without affecting fidelity by disbursing the workflow to multiple reasoning parsers.

[0016]    {0021]-Once the initial Level 1 pass is complete, the EF Feedback Loop and Utility Function allow the user to set the number and/or relevance scale to the first order of Level 2. The system will automatically push the most relevant sources to Level 2 so as to allow that portion of the system to apply its independent "noun phrase" parsing and "co-occurrence" algorithms to the classification/ categorization process. The Level 2 processor will then push only its new classification/categorization concepts back to level 1 for re-indexing. Following the second pass the EF FeedBack Loop and its associated Utility functions allows the second pass to Level 2 to take its most relevant data to Level 3 for its independent "verb" parsing algorithms. New concepts or classifications are passed back from Level 2 and to Level 1 for re-indexing and with results returned to Level 2. The EF Feedback Loop has now allowed 5 sets of algorithms to apply 3 independent sets of metadata markings that are all read in their entirety, in exactly the same fashion by the integrated system prior to the user seeing the first query result.

[0017]    {0022]-The EF Feedback Loop is controlled by a set of "Utility Functions" which are designed to support the centralization of information technology services that are of common concern to the Intelligence Community. This methodology employs the indexing schema in the same manner for structured and unstructured data, however we employ the specific use of structured data OLAP tools to address the EF FeedBack Loop independently from the noun phrase or verb parsing.

Brief Description of Drawings

[0023] ~~FIG. 1 Illustrates the challenge of scalability, which shows how very large data corpora must be processed in order for to extract meaning relative to a given inquiry.~~

[0024] ~~FIG. 2 is exemplary schematic views of the seven levels for a complete KD architecture includes five representation levels (1 through 5) and two control levels (6 and 7), in accordance with the invented method and apparatus. This figure shows the EagleForce "Representation Levels" concept, which is a foundation for building a knowledge discovery architecture. Levels 1 through 5 are detailed with Level 0 indexing (not shown) being reserved for the ingestion of extremely large data sets. Level 6 provides feedback control of lower levels, and Level 7 contains a utility function that is used to optimize feedback. This scalability serves to significantly enrich the metatagging process.~~

[0025] ~~FIG. 3 is provides a schematic view of data flow through the apparatus, including the optional step 0, but not reflecting optional step 5c, beginning with the original data corpus and the transformation of the data corpus through the operations performed upon the data corpus.~~

Detailed Description

[0026] ~~DESCRIPTION OF THE ARCHITECTURE~~

[0018] ~~[0027]~~ The method and apparatus consists of a tiered set of representation levels, herein described as five representation levels, along with an optional Level 0, together with the EF FeedBack Loop methodology and the EF Utility Function, which is designed to index, classify, and categorize data at eight levels of processing. The preferred embodiment is to employ a COTS-based architecture, making use of "best of the breed" existing and proven tools.

[0019] ~~[0028] existing and proven tools.~~ This embodiment has, in cooperation with several COTS vendors, developed and already demonstrated an integrated architecture with essential capabilities from Levels 1 through4. he addition of the technology provided by a Level 5 capability will complete the basic suite. Note that within this architectural framework, there is typically more than one COTS capability. Within the overall architectural concept, it is possible to use a

customer-preference for a specific COTS product within a given appropriate level, or to use more than one COTS capability, again within a given level.

[0020]    [0029] The EF FeedBack Loop begins with the order of scalability assuming that the incoming data set is on the order of 1 terabyte. The first order of business is to determine the time interval (Day, Month) to provide a consistent measurement basis for evaluation. The approach allows the first order of indexing (identification of documents with key words) to be metatagged as they are found in the document without the generalization into classes, concepts, co-occurrence-, etc. This level is used as the heavy lift, which allows the system and not the user to initiate the definition process as to whether a document has any potential relevance whatsoever, or if it can just be tossed. The goal at Level 0 is to reduce the amount of data as much as possible, without losing anything potentially useful.

[0021]    [0030] The preferred embodiment for this method and apparatus is based on a "Plug and Play" mindset. Thus, both the method and the apparatus are agnostic with respect to database vendor. A similar approach is employed throughout the architecture for the apparatus.

[0031]            INTERFACE DESCRIPTION WITHIN THE ARCHITECTURE

[0022]    [0032] There are two different classes of interfaces within the architecture. The first, and generally more straightforward, is the passing of data and metadata between tools. This apparatus and method solves the associated interface problems between several different tools, usually by a combination of special interface code at the API level, and use of intelligence in tool-specific metadata. Additional tools can be integrated as necessary.

[0023]    [0033] The second interface type involves passing of control between applications. This method and architecture has solved this via the EF Feedback Loop and the EF Utility Functions. The EF Feedback Loop has been described

in the said claim (6). The EF Utility Functions are a set of measures of the value (utility) of an intermediate or final output to the end-user, and have been described in the said claim (7). Utility functions thus provide a metric by which a proposed feedback action can be measured, and the overall performance of the system improved. Multiple utility functions are typically required because there are several independent axes that may be used to determine effectiveness.

[0034]        ~~ADVANTAGES AND BENEFITS OF THE METHOD AND APPARATUS~~

[0024]   [0035] This method and apparatus provide multiple benefits to the end user. Since the architecture comprehends the value of common look and feel, the usual difficulties in switching from tool to tool are mitigated. As capability is added, an increasing number of queries can be formed in natural language (English). In addition to facilitating ease of use and productivity, both of these factors reduce the amount of training required to employ these capabilities. Addition of a vector-based geo-referencing capability will enable the user to "drill down" based on geospatial locality.

[0025]   [0036] Advantageously, the invented apparatus and method can be used to preferentially extract relatively sparse concept classes and most especially various combinations of concept classes (where each "concept class" can be expressed as a category, a set of nouns and / or noun phrases, or a single noun or noun phrase, depending on the embodiment of the invention) along with identification of the relationships (single or multiple verbs, or verb sets) linking different concept classes. At the same time, the influence of "contextual" information can be incorporated to preferentially refine a given concept class, or to add more information relative to an area of inquiry. As an example, including geo-spatial references at Level 4 of the processing allows for "neighborhoods" surrounding a given occurrence to be preferentially tagged via feedback into the Level 1 process. Similarly, use of a Language Variant method at Level 4 can be used to identify geospatial regions of interest when a name of interest (found

during Level 1 or Level 2 processing) is identified and then one or more Language Variants of that name are identified in Level 4. If occurrences of these proper name Language Variants are then found as a result of feedback into a lower level (e.g., Level 1), then the geospatially-referenced regions associated with the Language Variants provide context for later iterations of the feed forward process that begins at Level 1.

[0026]  [0037] These together with other features and advantages, which will become subsequently apparent, reside in the details of construction and operation of the invented apparatus and method as more fully hereinafter described and claimed, reference being made to the accompanying drawing, forming a part hereof, wherein like numerals refer to like parts throughout the view.

~~Claims~~

~~[c1]~~

[0027]  ~~A~~An embodiment of the present invention includes a method comprising the steps of:

a) Performing Level 1: Indexing / Classification applied to data corpus "A", where "A" is a data corpus consisting of (typically) a large to very large number of members which can be structured, semi-structured, and/or unstructured text, the result(s) of any form of speech-to-text conversion, and/or images or other signal-processed data, and/or any combination of such data, where the Indexing / Classification process is performed specifically as: indexing and /or classifying the members of data corpus "A" by appending to each member one or more "metatags" descriptive of the content of that member, whether that content is explicitly referenced (e.g., via "indexing," using methods and terminology well known to practitioners of the art), or implicitly referenced using one or more of the various possible "classification"

algorithms (e.g., Bayesian, or Bayesian augmented with "Shannon Information Theory" feature vector weighting), where the only specific requirement of the classification algorithm(s) at least one of the algorithm(s) employed be "controllable" through at least one parameter value (e.g., the "sigma" value in a Bayesian classifier, or more broadly, the "sigma" value, the number of elements in the prototyping "feature vector" for such a classifier, and the "feature vector element weights" applied to each element of a given "feature vector," where these terms and associated methods are all well known to practitioners of the art, and this specification of possible parameter types is by no means exhaustive), and the end result is the set of one or more "metatags" so produced by application of one or more classification algorithm(s) to a given data corpus element and then associated with that element are indicative of the content of each element; and additionally a document may be classified and / or metatagged as containing one or more concept classes whose existence is inferred through the presence of certain words (typically noted as feature vectors) in that document,

b) Performing Level 1 to Level 2 Transition, by which a proper subset of members from the initial data corpus "A" are selected for Level 2 processing, which is done by selecting from among all the (optionally indexed and) metatagged members of data corpus "A" those whose metatags are a match to a set of criteria, where these criteria can be set either or both by the user of this method or by an automated process incorporated as part of this method, and whose exact specification does not in any way impact the generality of the method described here, and this subset is denoted data corpus "B",

c) Performing Level 2 Pairwise Associative Processing, by which the data corpus "B" members selected during said step (b) are processed so as to produce "pairwise associations" between the elements of each of these members of "B", where a typical embodiment of this step would be to generate a set of pairwise associations of nouns and / or noun phrases extracted from a text-based corpus "B", although this method can be extended and applied to data corpora containing other types of elements (e.g. images, signals) without loss of meaning or generality, and where the associations are typically limited to those within a given member of "B", although the results of such associations are typically noted accumulatively across the entire corpus "B", and a typical embodiment of this step is a "pairwise co-occurrence matrix" applied to objects in each member of "B" whereby a corresponding matrix element is incremented whenever a given pair of nouns and / or noun phrases occurs within a set distance of each other, although any accumulative pairwise-association method applied across "B" may be used without loss of the generality or meaning of the knowledge discovery method being described herein,.

[c2]

[0028]    A method as claimed in c1 further comprisingAccording to another embodiment of the present invention, the method may include the optional stepsteps of:

a) Performing Level 0: Optional Preprocessing / Indexing, specifically: (optionally) indexing the members of a data corpus "A0" by "tagging" each member of the corpus with one or more "metatags" in any such manner as is well known to practitioners of the art, whereby the "metatags" refer to specific identifiable elements (e.g., but not limited to, specific words, or specific content as might be found in an image) and where this step is typically reserved for very large data corpora (e.g.,

typically where the number of members of data corpus "A0" exceeds O(106)) but may be applied to any size corpus without loss of the validity or generality of this method;

b) Performing Level 0 to Level 1 Transition, specifically selecting those members of the data corpus whose "indices" as found and applied in said step (a) are a "match" to some specified criteria, whether these criteria are set manually by a user for a given knowledge discovery task or set via an automated process, and the method by which these "index matches" are selected is any one of those well known to practitioners of the art and detailed specification of such method or development of a new "indexing" method is not essential to specifying this knowledge discovery method, nor is it essential to specify the method by which such "indexed" data corpus members are "selected" for "Transition" to the predecessor step (1a) except that the general intention of said "selection" is to reduce the size of the "selected" sub-corpus, which we now denote corpus "A".

~~[c3]~~

~~A method as claimed in either c1 or c2, further comprising the step of:~~
[0029] According to another embodiment of the present invention, the method may include the steps of:

a) Performing Level 2 to Level 3 Transition, by which the "pairwise associations" found in said step (1c) are filtered by any one or more of various algorithmic means well known to the practitioners of this art so as to extract a subset of associations by application of one or more selection criteria, and the generality and meaning of this method is not dependent upon the specific nature of these criteria, and where a typical embodiment of this method would be to use a cut-off process selecting

only those "pairwise associations" that reach a certain predefined or preset value, whether this value is fixed or determined by an algorithmic means (such as histogramming or thresholding, or any such method as is employed by the community for similar purposes), and where extracted subset of these associations is hereafter referred to as data corpus "C" and is passed to a subsequent "Level 3" for further processing,

b) Performing Level 3 Syntactic Associative Processing, by which the data corpus "C" members selected during said step (3a) are processed so as to produce "syntactic associations" between the elements of one or more of each of these members of "C", where a typical embodiment of this step would be to generate a set of subject noun-verb-object noun associations using nouns and / or noun phrases extracted from the data corpus "C" as subject nouns (and potentially also as object nouns) and the verbs and additional object nouns are drawn from the data sources from which data corpus "B" was extracted, although this method can also include simple subject noun-verb associations and also verb-object noun associations, and where the identifications of subject nouns, object nouns, noun phrases, concept classes, and verbs, are those common to practitioners of the art, and the resulting representation of the syntactically-associated may be either in structured (e.g., database) or other form, so long as the syntactic relationship between the associated words or phrases is represented, and may also include, without loss of generality or meaning of this method, additional grammatical annotations to the basic syntactic representation (e.g., adjectives, etc.) and any one or more noun and / or noun phrase may be replaced with an associated "concept class, "using methods that are the same or similar to those described in (1a).

[c4]

~~A method as claimed in c3, further comprising the step of:~~
[0030]     <u>In yet another alternative embodiment of the present invention, the may
includes the steps of:</u>

a) Performing Level 3 to Level 4 Transition, by which the "syntactic
associations" found in said step (3b) are filtered by any one or more of
various algorithmic means well known to the practitioners of this art so
as to extract a subset of associations by application of one or more
selection criteria, and the generality and meaning of this method is not
dependent upon the specific nature of these criteria, and this subset
denoted as data corpus "D" is passed to Level 4 for further processing,

b) Performing Level 4 Context-Based Processing, by which the data
corpus "D" members selected during said step (4a) are processed so as to
produce "context associations" using one or more of a variety of
methods, which may be applied to either or both the elements of data
corpus "D" or to additional databases and / or knowledge sources, such
as are known to practitioners of the art, so as to extract refinement of
both associations and concept classes as was described in said step
(1a)<del>,</del><ins>.</ins>

[c5]

[0031]     <del>A</del><ins>The</ins> method <del>as claimed in c4, further comprising the step</del><ins>may also include the
steps</ins> of:

a) Performing Level 4 to Level 5 Transition, by which the "context
associations" and / or context refinements found in said step (4b) are
filtered by any one or more of various algorithmic means well known to
the practitioners of this art so as to extract a subset of associations by
application of one or more selection criteria, and the generality and

meaning of this method is not dependent upon the specific nature of these criteria, and this subset denoted as data corpus "E" is passed to Level 5 for further processing,

b) Performing Level 5 Semantic-Based Processing, by which the data corpus "E" members selected during said step (5a) are processed so as to produce "semantic associations" and "semantic meaning and / or interpretation" using one or more of a variety of methods, such as are known to practitioners of the art, so as to extract further refinement of associations as was described in said steps (2b, 3b, and 4b), concept classes as was described in said step (1a), and additionally any knowledge-based and / or semantic-based information that can be associated with the elements of data corpus "E",

c) (Optionally) perform steps 5a and 5b as many times as necessary with defined processing unique to each step 5c and different from any previous step to define the apparatus to the number of levels desired.


~~[c6]~~

~~A method as claimed in c1, c2, c3, c4 and / or c5, further comprising the step of:~~
[0032]  ~~Performing~~According to yet another alternative of the present invention, the above mentioned methods may also include the step of: performing Level N to Level (N-X) Feedback Control, where "N" errors to any of Levels 2 through 5, and "X" may take on any value from (1, ..., N-1) inclusive, by which one or more of the parameters governing any of the processes as described ~~in said claims 1, 2, 3, and / or 4~~above are controlled by the feedback loop operating on outputs computed at Level N, where N > the controlled level (1, 2, 3, or 4), and where multiple feedback loops can be implemented in any given instantiation of this method~~,~~.

[c7]

[0033] ~~method as claimed in c6, further comprising~~<u>According to another alternative embodiment, the method may also include</u> the step of: ~~Performing~~ <u>performing</u> a Utility Function computation and output, by which the "Feedback Loop" as described in said step (6) is modulated and controlled by means of a function so as to give either or both the user and / or an automated process the ability to control and "tune" the feedback loop so as to bring the overall system results to a desired level of performance, and where the formulation of said "Utility Function" follows he rules of practice as are well understood by practitioners of the art~~,~~.

[c8]

[0034] ~~An~~<u>The embodiments of the present invention also include an</u> apparatus for use with the processes described ~~in said c1, the apparatus comprising~~<u>above and including</u>: one or more data access and / or storage unit(s) "DS-1" coupled to receive and store as needed the data corpus "A", one or more computational processing unit(s) "CPU-1" coupled o receive the data corpus "A" and perform the processing as indicated in claim 1 "Level 1" processing, one or more data storage unit(s) "DS-2" coupled to the computational processing unit "CPU-1" so as to receive and store the data corpus "B" that is generated as an output of the process described ~~in said claim 1~~<u>above as</u> "Level 1" processing~~,~~.

[0035] <u>The apparatus may also include</u> one or more computational processing unit(s) "CPU-2" coupled to receive the data corpus "B" from "DS-2" and perform the processing <u>described above</u> as ~~indicated in claim 1~~ "Level 2" processing~~,~~.

[0036] <u>Furthermore, the apparatus may also include</u> one or more data storage unit(s) "DS-3" coupled to the computational processing unit "CPU-2" so as to receive and store the data corpus "C" that is generated as an output of the process described ~~in said claim 1~~<u>above as</u> "Level 2" processing~~,~~.

[0037]    The apparatus may also include a visualization and / or display unit or other means of providing viewing and / or results interpretation of either or both Level 1 and / or Level 2 processing, and / or making these results available to another process, whether automated and /or semi-automated~.

[c9]

[0038]    AnAccording to yet another embodiment the apparatus as claimed in c8, wherein if said step (2) is employed as part of the method, then additionally there is:, may one or more data access and / or storage unit(s) "DS-0" coupled to receive and store as needed the data corpus "A0", from stored and / or live data feeds, one or more computational processing unit(s) "CPU-0" coupled to receive the data corpus "A0" and perform the processing as indicated in claim 2 "Level 0" processing, and is for that purpose coupled to "DS-1" so that the outputs of the Level 0 processing can be stored and made available for Step (1)~.

[0039]    (optionally)Alternatively, a visualization and / or display unit or other means of providing viewing and / or results interpretation of Level 0 processing, and / or making these results available to another process, whether automated and /or semi-automated~ may be provided.

[c10]

An apparatus as claimed in c8, or c9, wherein if said step (3) is employed as part of the method, then additionally there is:
[0040]    When the apparatus employs Level 3 processing the apparatus may include one or more computational processing unit(s) "CPU-3" coupled to receive the data corpus "C" from "DS-3" and perform the processing as indicated in claim 3 "Level 3" processing, one or more data storage unit(s) "DS-4" coupled to the computational processing unit "CPU-3" so as to receive and store the data corpus "D" that is generated as an output of the process described in said claim

3 "Level 3" processing, (optionally)above as "Level 3" processing. In addition, the apparatus may include one or more visualization and / or display unit(s) or other means of providing viewing and / or results interpretation of Level 3 processing, and / or making these results available to another process, whether automated and /or semi-automated,.

[c11]

An apparatus as claimed in c10, wherein if said step (4) is employed as part of the method, then additionally there is:

[0041] one or more computational processing unit(s) "CPU-4" coupled to receive the data corpus "D" and perform the processing as indicated in claim 4 "According to another alternative embodiment of the present invention, when context based processing of level 4 is provided the apparatus may include one or more computational processing unit(s) "CPU-4" coupled to receive the data corpus "D" and perform the "Level 4" processing, and if more than one unit is so used, then appropriate coupling exists so as to transfer results between the processes as is necessary, one or more data storage unit(s) "DS-5" coupled to the computational processing unit "CPU-4" so as to receive and store the data corpus "E" thatis generated as an output of the process described in said claim 4 "Level 4" processing,.

[0042] (Furthermore, the apparatus may optionally) include one or more visualization and / or display unit(s) or other means of providing viewing and / or results interpretation of Level 4 processing, and / or making these results available to another process, whether automated and /or semi-automated,.

[c12]

An apparatus as claimed in c11, wherein if said step (5) is employed as part of the method, then additionally there is:

[0043]    According to yet another alternative embodiment of the present invention, the apparatus may include one or more computational processing unit(s) "CPU-5" coupled to receive the data corpus "E" and perform the processing as indicated in claim 5 "Level 5" processing, one or more data storage unit(s) "DS-6" coupled to the computational processing unit "CPU-5" so as to receive and store the data corpus "F" that is generated as an output of the process described in said claim 5above as "Level 5" processing, .

[0044]    (optionally)Furthermore, the apparatus may include a visualization and / or display unit or other means of providing viewing and / or results interpretation of Level 5 processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c13]

[0045]    An apparatus as claimed in c12,According to an exemplary embodiment of the present invention, an apparatus which additionally contains one or more computational and data storage units wherein the one or more "Feedback Loop(s)" as described in saidabove with regard ot step (6) are computed and stored, is provided.  The CPU and which is (storage units are) coupled to the appropriate Level N and Level (N-X) computational (CPU) units, (optionally) a visualization and / or display unit or other means of providing viewing and / or results interpretation of Feedback Loop processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c14]

[0046]    AnThe embodiments of the present invention also includes an apparatus as claimed in c13, which additionally contains one or more units wherein the one or more "Utility Function(s)" as described in said step (7) are computed, and

which is (are) coupled to the appropriate "Feedback Loop" computational (CPU) units~~,~~.

[0047]    ~~(optionally) a~~ A visualization and / or display unit or other means of providing viewing and / or results interpretation of the one or more Utility Function(s), and / or making these results available to another process, whether automated and /or semi-automated~~,~~ may optionally be provided.

~~[c15]~~

[0048]    An apparatus ~~as claimed in c14,~~ wherein the various units described in Claims (8) through (13) ~~inclusive~~above may be combined as appropriate for the purpose of enabling the processing and storage requirements ~~as are needed to meet the stated~~ ~~purposes of Claims (1) through (7)~~.

~~[c16]~~

[0049]    ~~An apparatus as claimed in c15,~~According to exemplary embodiments of the present invention, the apparatus described above may include, wherein one or more of the various units and the processes which are supported by each unit or appropriate combination of data storage and computational processing units, is embodied as an existing tool, whether available as a research prototype or "commercial-off-the-shelf" implementation.

~~Abstract of Disclosure~~

# ABSTRACT

[0038] The invented apparatus performs "A knowledge discovery "by extractingapparatus and method that extracts both specifically desired as well as pertinent and relevant information to query from a corpus of multiple elements that can be structured (elements are ordered according to some schema that defines type of element and length of element), unstructured (elements are undefined as to type and length and usually embedded in a document), unstructured, and/or semi-structured, along with imagery, video, speech, and other forms of data representation, to generate a set of outputs with a confidence metric applied to the match of the output against the query. The invented apparatus includes a multi-level (typically embodying Level 1 through Level 5, but nothing in the apparatus limits the multiple levels to 5) architecture, with an optional Level 0, along with one or more feedback loop(s) (Level 6) from any Level N to any lower Level (where N >=2), whereby the output of the lower Level can be controlled, along with a Level 7 utility function which governs the operation of the Level 6 feedback loop so that a user can control the output of this knowledge discovery method via providing inputs to the utility function.

Figures

ft.